OBSERVABLEND: APPLICATION OF OBSERVABLE LINGUISTIC FEATURES TO IMPROVE MACHINE LEARNING PREDICTIONS OF ENGLISH LEXICAL BLENDS

Jarem Saunders

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Linguistics in the College of Arts and Sciences.

Chapel Hill
2023

Approved by:

Katya Pertsova

Elliott Moreton

Jennifer Smith

ABSTRACT

Jarem Saunders: OBSERVABLEND: Application of Observable Linguistic Features to Improve Machine Learning Predictions of English Lexical Blends
(Under the direction of Katya Pertsova)

The process of lexical blending is a widely attested cross-linguistic process of generating new lexical items by combining two or more existing words. Despite its ubiquity, the structure of a blend is difficult to reliably predict, even when the order of the constituent words is known. This difficulty has been shown by machine learning approaches in blend modeling, including attempts using then state-of-the-art LSTM deep neural networks trained on character embeddings, which were able to predict lexical blends given the ordered constituent words in less than half of cases, in the best performing models.

This project introduces a novel model architecture which demonstrates notable increases in the rates of correctly predicted lexical blends using variations on Logistic regression and Random Forest learners. This is achieved by generating multiple possible blend candidates for each input word pairing and evaluating them based on observable linguistic features. The feature system in question is also manipulated, demonstrating that models trained on phonologically-determined observable features outperform those trained using purely orthographically-derived feature sets. The success of this model architecture illustrates the potential usefulness of observable linguistic features for problems that elude more advanced models which utilize only features discovered in latent space, and lays the groundwork for a more linguistically-motivated and interpretable approach to the generation of English lexical blends.

To my wonderful wife, Lily, whose love and support made this work possible.

Thank you for always believing in me and in this project.

ACKNOWLEDGEMENTS

I wish to express my gratitude to my thesis advisor, Dr. Katya Pertsova, for continually offering guidance on this project and helping me to crystalize the direction of the research. The other members of my thesis advisory committee, Dr. Jen Smith and Dr. Elliott Moreton, also offered invaluable advice and perspective in conducting this research. I also could not have completed this work without the feedback and encouragement I received from other faculty members and my fellow grad students in the Department of Linguistics, especially the members of the P-Side research group who allowed me to share about this project on several occasions. To each of these contributors, and to all of my non-linguist family members and friends who stood by me through this work—and surely heard more than you ever wanted to know about lexical blending in the process—thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: PREDICTIVE MODELS OF LEXICAL BLENDING

Section 1.1: Problem Statement

This project was designed to examine the feasibility of using observable linguistic features in a modified classification task to predict lexical blends of English, given two ordered constituent words. Several classification learners and feature set variations were tested to determine which, if any, of the possible combinations would be able to outperform previous benchmarks in performance for machine learning approaches to blend modeling.

Section 1.2: Lexical Blend Background

Lexical blending is a linguistic phenomenon attested in numerous languages of the world. Despite its ubiquity, it is often difficult to predict exactly how two words will be combined in a given language. English blends are most typically created by combining segments of two words such that some initial portion of the first word and some final portion of the second word is preserved, and blends of this type constitute 90% of blends of English or higher, depending on the blend corpus. Even within this group, there is considerable variation in the proportion of the constituent words which are retained in the blend and where the boundary between the constituent words is found.

A number of constraints have been identified that are related to blend formation, though these features are rarely found to apply for all instances of blend generation. Rather, these constraints express general trends in blend formation which frequently appear to be at least partially violable and cannot account for all observed blend structures, even in the most common concatenation patterns.

Predicting blends has been described, in large part, as a matter of determining the position in each word at which it is truncated, known as the switchpoint. Blends tend to preserve metrical structure of the second word, and especially tend to keep the primary stress in the same position as the second input word (Arndt-Lappe & Plag 2013). This often means that the switchpoint occurs before the position of primary stress, thus preserving the entire rime of the primary stress syllable, but even if the segmental content changes, the stress frequently remains in the same position.

It has also been noted that switchpoint tends to occur no more than 1 syllable away from the second word's primary stress syllable (Gries 2004). The length of the input words has long been known to be influential in deciding the switchpoint, though there is no clear agreement regarding its exact effect. Bat-El (2006) claims that blends tend to match the length of the second input word while Bauer describes a tendency for the length of the second input word to be the maximum length of the blend. Cannon (1986) and Gries (2004) find that the first input word also may have an effect, and make the claim that blends tend to match the length of the longer of the two input words.

Switchpoint placement seems also to be influenced by the syllabic structure, usually occurring at boundaries between syllabic constituents (Gries 2012, Kelly 2009). There also seems to be a tendency for the switchpoint to occur at onset boundaries in the first word, rather than at coda boundaries (Gries 2004, 2006). Despite the fact that these tendencies have been well documented in the literature for many years, no efforts have been made to date to utilize these features in a unified model to predict blend switchpoints.

Section 1.3: Research on Machine Blend Prediction

By investigating any of the numerous 'portmanteau generators' on the internet that operate on simple, rule-based substring concatenation, it becomes readily apparent that, while some blends follow easily predictable patterns, there are many others which cannot be generated so easily. These generators show that it is difficult to find generalizations which hold for large numbers of blends, as evidenced by that fact that when tasked with recreating even attested blends, these generators often formulate phonotactically and orthographically invalid words of English.

A small number of research projects have attempted to use machine learning methods to produce lexical blends, and have had, overall, quite moderate success in predicting blends. Trained on relatively small language corpora, these projects have struggled to predict blends with high accuracy, and incorrect predictions often suffer from the same issue of producing substring concatenations that are highly marked or even impossible structures in English. The two most serious attempts using machine learning to predict blend structure are Deri & Knight (2015) and Gangal et. al. (2017).

1.3.1: FST Learners for Blend Generation

The earlier of these attempts, Deri & Knight (2015), employed a multi-tape FST framework and a system of grapheme-phoneme alignment to generate blends, and met with some success. By training on the aligned sequences of phonemes and graphemes, the model was intended to learn the segment by segment transformations needed to produce a blend from two input words. It was able to correctly predict a blend in nearly 50% of cases, given the input words. However, the incorrect predictions often did not contain recognizable portions of the input words or were highly phonotactically marked.

1.3.2 LSTMs for Blend Generation

Similar levels of success were found in Gangal et. al. (2017), which used recurrent neural networks with character embeddings trained on English text to predict blends based on orthography alone. Using the LSTM architecture, the model utilized an encoder/decoder trained on a large amount of English text in order to approximate English orthotactics and a recurrent neural network that trained on the encoded representations of blend components to arrive at the blend predictions. The study proposed several different models using a recurrent neural network architecture, but, like Deri & Knight, none of models was successful above 50% and all models had phonotactically implausible output forms in many of the incorrect output forms.

An important fact demonstrated by this latter study is that the problem of blend generation may be more effectively considered as a discriminative task, rather than a generative one, in the computational sense. They show that higher accuracy can be achieved by generating all possible sequential substring concatenations of the two input words and selecting the token with the highest probability of being a valid word of English, rather than trying to predict a blend output character by character. This was shown by comparing a greedy decoding strategy to one that generated all potentially blend-like concatenations of the two words and selected the one with the maximum probability. The discriminative approach was found to perform with higher rates of correct predictions in most model variations tested and ultimately yielded the highest rate of correct blend predictions.

Using this strategy, the problem is effectively reduced to the problem of finding the partition point in each of the two input forms that yields the highest likelihood of validity as an English lexical blend, rather than trying to generate a novel word of unknown length character by character. This method was successful in predicting the blend output for 48.8% of blends in

Gangal et al.'s models, compared with a maximum of 28.0% using character by character decoding.

The overall modest accuracy in predicting blends using any of the models used by Gangal et al. is notable because it shows a failure of a model architecture which has been used successfully for a wide array of natural language domains. This shows that blends present a unique challenge when compared to other sequence-to-sequence natural language processing tasks like morphological inflection or machine translation. (Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., & Hulden, M. 2016, Zoph, B., & Knight, K. 2016)

Section 1.4: Improving Predictions Using Domain-Specific Knowledge

A modification to the Gangal et al model was described by Kulkarni & Wang (2018) demonstrated that imposing limitations on the length of blends and the possible characters used during the encoding of character sequences to better reflect actual blend behavior led to an increase in performance for the Gangal et al model architecture. This was achieved using a more restricted embedding process for characters, which did not rely on the encoder-decoder architecture employed by Gangal et al. The improvements on the forward model were notable, but failed to break the benchmark for backward architecture, which produced the best performing model iteration of all those described in the original Gangal et al. paper. Nevertheless, the improvements obtained through this linguistically-motivated modification to the model demonstrate the potential usefulness of domain-specific knowledge in improving machine learning approaches to blend formation.

Section 1.5: Shortcomings of Previous Approaches

This paper seeks to further improve on existing approaches of blend prediction by making more significant changes to the feature space utilized during training. Whereas the models outlined in both Deri & Knight (2015) and Gangal et. al (2016) relied on models which trained on features learned in the latent space, this paper proposes a model that is designed to use a minimal number of pre-selected features that are believed to be relevant to blend formation based on existing linguistic research on lexical blends. Such models have been used in natural language processing tasks for decades, though they have decreased in popularity as more data-rich neural models without pre-tagged features have shown increased effectiveness in many different natural language processing tasks. Due to the unique shortcoming of such models on blends, especially the ungrammatical predictions they make, this paper moves away from state-of-the-art neural approaches with features learned in the latent space and revisits more classical machine learning methods. The architecture proposed hereafter utilizes observable features determined by linguistically-motivated feature preprocessing to investigate whether manually-selected features based on domain-specific knowledge can allow a simplified machine learning component to better address a problem that confounds more advanced architectures.

CHAPTER 2: METHODS

Section 2.1: Model Overview

The blend prediction system proposed in this thesis is designed to utilize known linguistic features of lexical blends, as well as explicit scoring of English phonology to improve upon the state-of-the-art prediction accuracy for English blends.

The model consists of the following 3 components:

1. A candidate generation component that produces possible blends for the model to consider, along with model features and labels for each candidate
2. A machine-learning component which applies weights to features and calculates probabilities for the candidates
3. A selection component which identifies the blend candidate that is most probable for each input sequence

2.1.1: The "Exhaustive Generation" Strategy

The underlying principle of this model is based on Gangal et al.'s "exhaustive generation" strategy, which creates possible blend substring concatenations of substrings that begin with the start of the first word and end in the final segment of the second word. In this way, we can instead consider the probabilities of the various possible concatenations and can select a candidate from a finite set, rather than generating segment by segment or character by character. Thus, using exhaustive generation to create the full candidate space lets us implement a model such that it directly learns to estimate the probability that a $\boldsymbol{P(Y_i = 1|X)}$, where $Y_i =$ describes the probability that a model instance Y_i is a valid lexical blend of English, given the set of features.

This system is comparable to a Maximum Entropy Model of phonology or related Harmonic Grammar models which allow for weighted constraint values. We can consider the generation strategy as analogous to the generator component of constraint-based linguistic models. The model's features function like weighted constraints of a Maximum Entropy model. Both are predicated on the assumption of some system that can both weigh constraints and also select a candidate of maximum desirability based on the learned weights.

2.1.2 The Learning Component

Having generated the set of candidates and their respective feature values and training labels, the model will use a machine learning component to determine which candidates are the best possible lexical blends. In this analysis, we entertain two distinct learners as the trainable component of the model, namely, logistic regression and random forests. Logistic regression is used as the primary method of feature selection, and is also considered using feature augmentation as the final predictive model. Random forest is entertained as an alternative to augmented logistic regression and the relative strengths and weaknesses of each are compared. Both of the model types entertained output probability values for each blend candidate which are used to select the optimal candidate in each blend set.

2.1.3 Candidate Selection and Evaluation

Using the probabilities generated by the learner, the selection component chooses exactly 1 winning candidate for each input pair. The primary model comparison metric is the percentage of blends for which the model correctly predicts the winning output. This metric is standard to the previous two blend modeling approaches.

In addition to prediction rate, each model and feature set pairing will be evaluated on Levenshtein edit distance between the predicted output and the correct blend form. This metric shows the number of character edits (additions, deletions, and substitutions) required to make two strings identical. It was also used by both Deri & Knight and Gangal et. al, and will therefore allow the performance of this model to be more directly compared to theirs.

The final metric used will be the number of blends for which the correct candidate is among the three candidates with the highest probability assigned by the model. This metric was not used in previous blend modeling approaches, but is introduced here to account for the fact that, for naturally occurring lexical blends, there are often many plausible candidates when the blend is initially coined (Arndt-Lappe & Plag). Ultimately, one form will win out, but before it has been lexicalized there will often be alternatives. It is also noted by Gries (2006) that the lexicalized form of blends is not always the blend candidate that quantitative analysis would predict to be the most likely. For this reason, we consider the 3 candidates to which the model assigns highest probability of being a good blend candidate to account for the possibility that the model has given a high probability to the correct candidate, but ultimately selected an alternative blend as the winner.

2.1.4 Cross Validation

Each data analysis metric will be evaluated using 10-fold cross validation, and the average prediction rate will be reported over the 10 folds. In this procedure (k-fold cross validation, generally), training data is randomly partitioned into 10 subsamples, or folds, of the data set, and the model is iteratively trained 10 times, each time with a different fold acting as the testing data and remaining 9 partitions serving as training data. In this way, the model's performance is evaluated over 10 different 90/10 training/test splits, and each instance in the training set is used as part of the test set for exactly one training iteration. For this model, it is used both to better assess the generalizability of the results and to create randomized sampling distributions from the performance of the various models and feature sets in order to statistically compare their performance.

Section 2.2: Language Corpora Used

The model architecture described in this project was applied to model iterations that were trained and tested using 3 different corpora. One of these had been used for empirical study of blends but not blend prediction/generation, and the two others were those used in the prior attempts at machine learning for blend prediction.

2.2.1 Shaw (2014) Corpus

The primary corpus used for development and testing for this project was the corpus developed by Shaw (2014) for investigation of positional faithfulness effects in blend formation, and is a curated set of blends from a dictionary of blends by Thurner (1993). This corpus was selected both for its size and for its linguistically-verified data quality. It contains 1,395 blend forms in total and has been used previously in successful studies of phonological patterns of blending.

As noted above, this project is only concerned with blend forms which preserve some initial portion of the first input word and some final portion of the second word, so other types of blends, most notably embedding blends and blends of more than two words, were excluded from the training set. This removed 189 items from the total count. Other blends were excluded because phonetic information could not be found for one or both of their respective input words, resulting in 116 additional corpus items that were not considered in the training set. In total, 1096 blends remained to make up the training data for the project.

2.2.2 Gangal et al. (2017) Corpus

The second corpus used is the corpus produced by Gangal et al. for their paper *Charmanteau* (2017). Although this corpus is larger than the Shaw corpus at 1624 entries, the sourcing for most of the blends is somewhat unclear. Gangal et al. cite the following as sources for their blend data, but they give no indication as to the sourcing of specific corpus items, nor do they state the proportion of entries contributed by each source, instead simply stating that they were retrieved from "Urban Dictionary, Wikipedia, Wiktionary, BCU's Neologisms List from '92 to '12."

Due to the fact that many of the blend sources noted in the corpus are easily editable and restricted to online use only, it is not known whether the blends are attested usages in a particular speech community, or simply one-time coinages that may or may not be representative of typical lexicalized blends.

This corpus also had a slightly higher proportion of input words that did not appear in the phonetic dictionaries, leading to 1,094 items from this corpus being used for training and testing.

The corpus also includes a large number of terms acknowledged by the authors to be potentially offensive. Given this fact, this corpus was used only as a control for comparing to the previous baseline, and not as a developmental guide for the model.

2.2.3 Deri & Knight (2015) Corpus

The final corpus is the smallest, containing exactly 400 words before filtering . This corpus was also not used in development, but rather preserved for use as a comparison point to assess the efficacy of the proposed model against the performance of Deri & Knight's model using the same training data. The source for this list is also credited to Wikipedia and

Wiktionary, but it is noted that the entries in this corpus were required to have some medial overlap and to fit other predefined patterns established by the researchers (Deri & Knight 2015).

After data attrition due to entries not found in all dictionary lookups needed and those whose final forms deviated from input orthography, a total of 322 items remained in this corpus for training and testing.

Section 2.3: Feature System

In total, 24 distinct training features were considered for the various iterations of the model. These come from several proposed explanations of the processes that guide English blend formation. These were included together in the feature set both to consider multiple different feature subsets to maximize the model's prediction accuracy and to provide insight into which features are most predictive of English lexical blending as a phonological process.

The faithfulness features used in the models are primarily based on the feature system laid out by Arndt-Lappe & Plag's 2013 paper *The role of prosodic structure in the formation of English blends*. Many of these features are prosodic indicators, which Arndt-Lappe & Plag found to be most predictive of English blend formation, but several others are based on alternative features proposed as constraints on blend formation. A small number of variables reported in other (mostly older) analyses of blends were included in the feature set, along with a feature designed to increase the phonotactic validity of output forms, namely scores generated by the BLICK phonotactic probability calculator, described below.

While it is noted elsewhere in this thesis that these model features can be considered as rough approximations of the faithfulness and markedness constraints of a MaxEnt model of phonology, it is important to note that this model is not intended to function as a psycholinguistically plausible model of blend formation. The model is not designed to approximate the precise mechanism used by humans acquiring English to learn to produce lexical blends, but it is intended to provide some insight into what cues speakers may be using to create blends based on the weighting of features utilized by the model.

2.3.1 Text-based Phonemic Features

Of the 24 features considered in this thesis, 23 are calculated based on properties of the phonemic representations of the input words. These phonemic representations are determined through a dictionary lookup using a pre-generated dictionary of grapheme-to-phoneme alignments which pair grapheme n-grams with the corresponding phonemes of the words as written in the CMU Pronouncing Dictionary. The alignments were generated using the *phonetisaurus* grapheme to phoneme library (Von Kleist 2015).

The set of features taken from Arndt-Lappe & Plag is given in Table 2.1, with features that measure identical values for each input word grouped together:

Feature names	Description	Number of distinct features values
initial length of Word 1/Word 2	number of syllables of each input word	2
length of candidate	number of syllables of the generated candidate	1
medial overlap	whether the prefix/suffix have overlapping phonemes	1
stress from right (Word 1, Word 2)	number of syllables from the right edge to syllable with primary stress	3
stress from left (Word 1, Word 2)	number of syllables from the left edge to syllable with primary stress	3
switchpoint Word 1	number of syllables from the left until the switchpoint	1
switchpoint Word 2	number of syllables from the right until the switchpoint	1
constituent boundary of switchpoint (Word 1, Word 2)	whether switchpoint in each word occurs onset, nucleus, or coda	6

Table 2.1: Phonologically-derived features adapted from Arndt-Lappe & Plag (2013)

In addition, these four features were calculated, based on usage in earlier descriptive analyses on blends. (Gries 2004, Kubozono 1990)

Feature names	Description	Number of distinct features values
Word 1/Word 2 primary stress survived	whether syllable bearing primary stress is preserved in each word	2
proportion of Word 1/Word 2 segments	proportion of segments from the input word retained in the blend candidate	2
proportion of Word 1/Word 2 syllables	proportion of syllable nuclei from the input word retained in the blend candidate	2
switchpoint at Word 2 primary stress syllable	whether the switchpoint in word 2 falls in the primary stress syllable	1

Table 2.2: Other phonologically-derived model features, adapted from Gries (2004) and Kubozono (1990)

2.3.2 BLICK Score of Phonotactic Markedness

The final feature calculated for each candidate was a measure of phonological markedness, represented using markedness score values of the BLICK phonotactic probability calculator. This tool uses the UCLA MaxEnt (Maximum Entropy) grammar learner to predict how likely a sequence of phonemes is as a word of English. Given a phoneme sequence as input, BLICK returns a markedness score representing how "good" an English word that sequence would make. This score is the sum of weighted constraint violations in the BLICK grammar for the phoneme sequence provided. The larger the score, the more improbable the sequence is a word of English, or, the more marked it is predicted to be.

The set of constraints in the BLICK model and their respective weightings are pre-trained from a corpus of high-frequency English words from the CMU pronouncing dictionary, with weighted constraints learned from n-grams of the phonetic features associated with the phoneme sequences of the words in the corpus (Hayes 2012). This pretraining is functionally similar to the word embeddings used by Gangal et al, but the way these constraints are learned is much more similar to the learning strategy of this proposed model, since the underlying mechanisms of the maximum entropy and logistic regression models are identical.

Section 2.4: Candidate Generation

2.4.1 Exhaustive Generation Using Alignments

The process of candidate generation involved isolating sub-strings from both words in each input word to concatenate together to create a possible lexical blend form from those words. For each segment in the phonemic representation of the first input word, blend candidate prefixes consisting of contiguous phonemes of the beginning of the word were iteratively produced. Similarly, a set of candidate suffixes were produced from contiguous phoneme sequences of the second input word ending in the final phoneme. Each prefix was combined with each suffix to produce the full candidate set for each word pairing. These candidate prefixes/suffixes were produced using grapheme/phoneme alignments, such that each prefix/suffix was paired with the corresponding orthographic representation of those phonemes from the input word.

candidate a

graphemes	b	l	u	n	c	h
phonemes	B	L	AH1	N	CH	

candidate b

graphemes	b	u	n	c	h
phonemes	B	AH1		CH	

candidate c

graphemes	b	n	c	h
phonemes	B	N	CH	

candidate d

graphemes	b	c	h
phonemes	B	CH	

candidate e

graphemes	b	r	l	u	n	c	h
phonemes	B	R	L	AH1	N	CH	

candidate f

graphemes	b	r	u	n	c	h
phonemes	B	R	AH1	N	CH	

….

Figure 2.1: Illustration of candidate generation process

This alignment process is necessary because it allows the model to train on the phonological properties of the word while still using the orthography of the blend produced in order to compare to the desired output blend for the purpose of establishing training/test labels for the data. Using this system to label the data allows us to consider blends for which there is not an accessible phonemic transcription of the desired output form, as the only way to establish the correct output for these words is the orthography. Due to the fact that many blends are novel coinages and do not appear in dictionaries, using orthography as the determining factor of labeling allows for a larger training set than using only instances in the dataset for which the desired blended form is given in its phonemic form.

Using the Shaw corpus, the average number of candidates produced for the input pairs is 30.78, with a minimum of 4 and a maximum of 149 candidates. With this finite, algorithmically-generated candidate space, the average base probability of selecting the correct candidate at random from the candidate set is roughly $\frac{1}{30}$. The average size of a candidate set for the Gangal et. al. and Deri & Knight corpora were similar, at 34.10 and 32.06 candidates per input word pair, respectively

2.4.2 Exclusions from the Candidate Set

The full candidate set for each blend consists of only the unique substring concatenations of the input words. For some word pairs, multiple substrings can produce identical candidates, as they will result in identical phonemic forms and therefore identical values for all model features. Such candidates were removed from the candidate set. Additionally, no candidates are included in the candidate set which preserve the entirety of the phonemes of both input words without overlap, as this process results in a compound, rather than a blend. Any candidate which reproduces the phonology or orthography of one of the input words, rendering it indistinguishable from that word, is also considered an invalid blend candidate and is excluded from the candidate set.

A small subset of correct blends in the corpus cannot be generated in this manner. This includes both so-called insertion blends, in which some or all of the second word is surrounded on both sides by the first word, words which preserve only initial portions from both constituent words, and a small number of concatenating blends which alter the orthography to better represent the phonemes of the resulting blend. In order to ensure that the correct blend exists for all candidate sets in the training data, these entries will be excluded from the training data input to the model.

Section 2.5: Feature Extraction

The model calculates values for all features for each candidate during the iterations of the candidate generation process, with the exception of BLICK scores, which are calculated in a separate step after the generation process was completed and the final phonological forms of each blend could be evaluated. The majority feature values were measured by simply counting the number of segments or syllables in either the input word or the candidate substrings at each level of the iterative process, while others required the use of syllable structure assignment values for each segment of the input words.

For each blend pairing, the following features are calculated from the phonemic representation of the input words, and are constant for all members of a given candidate set:

Distance from Word 1/Word 2 left edge to primary stress
Distance from Word 1/Word 2 right edge to primary stress
Word 1/Word 2 length

The following are properties of the separate candidate substrings generated from each word:

Distance from Word 1 left edge to switchpoint
Distance from Word 2 right edge to switchpoint
Switchpoint at Word 1 onset/coda/nucleus boundary
Switchpoint at Word 2 onset/coda/nucleus boundary
Switchpoint occurred at point of Word 2 primary stress
Word 1/Word 22 primary stress syllable survived

Finally, these features were dependent on the resultant blend when the substrings were combined:

Word 1/Word 2 proportion surviving syllables
Word 1/Word 21 proportion surviving segments
Candidate has medial overlap
Candidate length (syllables)

Model features referencing syllable position were calculated using the grapheme-phoneme alignments in conjunction with a dataset which gives the syllable segmentation points for each word. Given the sequence of phonemes and the syllable break points for each word, syllable constituent structure was assigned to each input segment based on its relationship to the syllable nuclei - segments appearing after a syllable boundary but before the nucleus were assigned onset position, those appearing after the nucleus were all assigned coda position.

The values for syllable structure boundary constraints could then be determined by looking for changes in syllable structure assignments on either side of the switchpoint for a blend candidate. If the phoneme on the left (Word 1) side of the switchpoint was the final phoneme in its source onset, nucleus, or coda in the input word, it was determined to be at a syllable boundary, and the corresponding Word 1 syllable boundary feature would be valued ‘1’ for the candidate. Similarly, if the phoneme on the right (Word 2) side of the switchpoint was the first phoneme in the onset, nucleus, or coda of source Word 2, the syllable boundary feature for that position was labeled ‘1’. If the phoneme bordering the switchpoint was not found to be at a syllable boundary, all 3 possible syllable position features for the source word of the phoneme were valued at 0.

2.5.1 Candidate Labeling

Once a candidate has been generated and its feature values determined, it is assigned a label of ‘1’ if it results in the desired blend output, and is labeled '0' if it does not. While the features extracted for model training are based on the phonological content of the input words, the label is determined by checking for a string match between the orthography of the generated candidate and the actual blend orthography. In this way, the grapheme portion of the grapheme/phoneme alignments truly serves only to provide labels for the training data and evaluate final predicted forms.

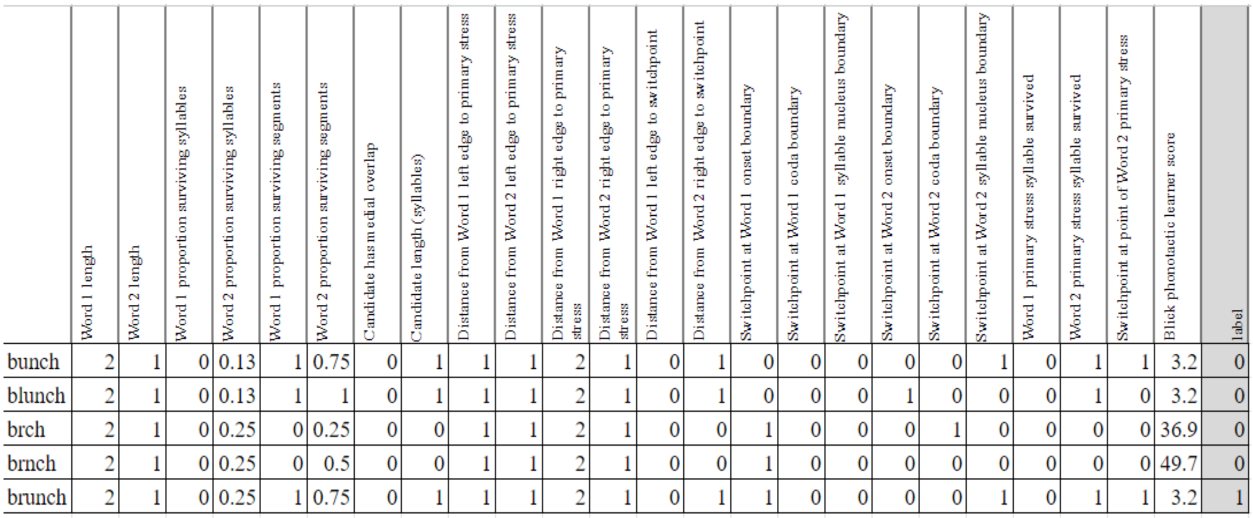
Labels are assigned such that for each candidate set there is exactly one candidate which is labeled as the correct blend output. For most blends, assigning a label requires simply checking whether a blend candidate has a grapheme that matches the desired orthographic output. However, because the generation process sometimes produces candidates which have identical phoneme sequences but differing orthography, there are times when two candidates match in all feature values, but one has orthography matching the output and one does not. In these instances, the generation algorithm enters only the candidate with the matching orthography into the candidate set, and the other is excluded.

This may seem initially to introduce bias into the model, as it is reducing the candidate set in order to boost the prior probability of finding a blend, but in truth is just treating the training data as if the phonetic representations for the blends were known and could be used as training labels, in which case there would be no conflict between these candidates.

A related scenario can occur when two candidates both have orthography which matches the desired blend output, but have differing phonology. This most commonly happens when the two candidates differ only in terms of a single vowel, typically when a reduced vowel shares its orthographic representation with a full vowel. In this case, rule-based heuristics are used to resolve the conflict and assign the positive training label to one of the candidates, but not the other. Specifically, the generation algorithm selects the candidate with a reduced vowel if the unreduced vowel in the other candidate is stressed and the candidate has a separate primary stress elsewhere. In cases in which the primary stressed syllable both or neither of the input words are present in the candidate, the generator will give preference to the candidate which preserves a greater portion of each of the input words. These heuristics have some potential for introducing bias into the interpretation of feature importance, but should not artificially inflate the performance of the model.

2.5.2 Feature Vectors

After feature extraction, each candidate has a total of 24 feature values, along with the training/test label associated with those features. This is represented in a vector of length 25. The candidate space for a given input word pairing can then be described as a matrix of size $N_c \times 25$, where N_c is the number of candidates in that set.



	Word 1 length	Word 2 length	Word 1 proportion surviving syllables	Word 2 proportion surviving syllables	Word 1 proportion surviving segments	Word 2 proportion surviving segments	Candidate has medial overlap	Candidate length (syllables)	Distance from Word 1 left edge to primary stress	Distance from Word 2 left edge to primary stress	Distance from Word 1 right edge to primary stress	Distance from Word 2 right edge to primary stress	Distance from Word 1 left edge to switchpoint	Distance from Word 2 right edge to switchpoint	Switchpoint at Word 1 onset boundary	Switchpoint at Word 1 coda boundary	Switchpoint at Word 1 syllable nucleus boundary	Switchpoint at Word 2 onset boundary	Switchpoint at Word 2 coda boundary	Switchpoint at Word 2 syllable nucleus boundary	Word 1 primary stress syllable survived	Word 2 primary stress syllable survived	Switchpoint at point of Word 2 primary stress	Blick phonotactic learner score	label
bunch	2	1	0	0.13	1	0.75	0	1	1	1	2	1	0	1	0	0	0	0	0	1	0	1	1	3.2	0
blunch	2	1	0	0.13	1	1	0	1	1	1	2	1	0	1	0	0	0	1	0	0	0	1	0	3.2	0
brch	2	1	0	0.25	0	0.25	0	0	1	1	2	1	0	0	1	0	0	0	1	0	0	0	0	36.9	0
brnch	2	1	0	0.25	0	0.5	0	0	1	1	2	1	0	0	1	0	0	0	0	0	0	0	0	49.7	0
brunch	2	1	0	0.25	1	0.75	0	1	1	1	2	1	0	1	1	0	0	0	0	1	0	1	1	3.2	1

Figure 2.2: Selection of Calculated Feature Values for Candidates. From Input Word Pair ‘breakfast’ + ‘lunch’

Section 2.6: Learner Components

Given an array of features for each candidate, the model uses a trainable component which gives a probability that a given candidate is a blend. In total, four distinct models were tested for this thesis. Three were variations of logistic regression (LASSO, ridge, and polynomial regression) and the final model was a random forest classifier.

2.6.1 Logistic Regression Component

This model architecture is among the most well-known algorithms for classification problems. Given a vector of input features X, the learner fits weights to the input features such that the sum of weighted features can be used to calculate probability P that a training instance (candidate) with features X belongs to a given class. Using labeled training data, the learner fits coefficients to the variables of the data vector and an intercept term, given as:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

These values of β are fit to the distribution of feature values of the training data, X_{train} by minimizing the logistic loss function through gradient ascent. Therefore, given the sigmoid probability function, which can be used to find the probability of membership in class:

$$P(t) = \frac{1}{1 - e^{-t}}$$

We can represent the probability P_{blend} that an instance is a valid blend of English as follows:

$$P_{blend} = \frac{1}{1 - e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots)}}$$

Once the learner has establish values for each coefficient β, these can be used to predict probabilities P_{blend} of test instances given their feature values X_{test}. The values of β learned in training can also be examined to gain insight into the patterns learned by the model.

2.6.2 Logistic Regression Variations

The specific variations of logistic regression used in this paper are LASSO, Ridge, and Polynomial Regression. Ridge and LASSO are both techniques for regularizing the data, which is applied during model training to reduce overfitting by restraining complexity. This is accomplished by imposing restrictions on the model training process such as penalties for overly large constraints, which is especially important if there are collinearities in the feature matrix that otherwise might introduce unwanted noise into the feature weighting. LASSO regularization can also be used for feature selection, as it is capable of reducing feature weights to 0 (Tibshirani 1996).

The final model used was Polynomial regression, which uses feature augmentation to include 2nd degree polynomials formed from the feature set, as well as interaction terms between the features. This increases model complexity, with 325 features instead of 24, using the whole corpus. This increase is notable, but the model still has far fewer features than would be necessary to create most systems of word embeddings used in deep learning. Regularization was also applied to the Polynomial Regression learner, specifically the scikit-learn default LASSO regularization method.

2.6.3 Training Procedure & Parameters

For this thesis, coefficients are found using the *LIBLINEAR* solver, as implemented in the free scikit-learn Python library (Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J.). Feature sets were also standardized using the standard score measure in order to make the feature weight more interpretable, though this has no effect on the actual probabilities calculated by the learner, and therefore does not affect the predictions of the model.

2.6.4 Random Forest

Random forest is an ensemble method for machine learning, meaning that it leverages the predictions of multiple different models to make its final predictions. For a random forest, multiple decision tree classification models are generated, which assign instances in a training set to a class based on a tree graph which branches on logical conjunctions of features in the training set. Using random subsets of the dataset and its features, the model trains multiple different trees and assigns class probabilities to instances in the test set based on the proportion of trees which assign it to a given class. For the overall model architecture, this probability is used in an identical fashion to those produced by logistic regression.

Section 2.7: Candidate Selection Component

The candidate selection component is the most straightforward component of the model. Given the probabilities for each candidate as the output of the blend, the selection component compares the probabilities of all members of the candidate set for each blend and selects the candidate with the highest probability as the winning candidate, and predicts that blend form as the output for that given blend.

Thus for a candidate set C for an arbitrary blend, the selection criterion for the solver may be expressed such that the model selects as the output $\hat{y}$ the candidate with feature set X_i which gives the maximum P_{blend} of any feature set of candidates found within the candidate set C for a given input word pairing.

$$\hat{y} = \ argmax\, P_{blend}(X_i),\ X_i\ \in\ C$$

If we again draw analogy to constraint-based linguistic models, the candidate selection component functions similarly to the evaluator component. In the same way that the evaluator in, or a Harmonic Grammar model choses a surface form based on candidates with lowest sum of weighted constraint violations and disregards all other potential surface forms, the selection component selects the blend candidate from each candidate set which has the maximum probability of being a valid English blend, given the weighted feature values of the system.

Section 2.8: Orthographic Proxy Models

In order to assess whether the phonological features add meaningfully to this model architecture, an alternative, orthographically-derived feature set was produced for use in the proposed models. These additional feature sets were extracted for each model using only features that can be obtained through surface-level orthographic properties of the blends, and function as proxies for the phonological features used in the primary model. These orthographic features were then used in training the same model types as the phonological features to provide a point of comparison for the phonological model, thereby providing an indication as to whether the model's performance is due to the features or purely to the model architecture.

Instead of counting syllables and segments, this feature set counts only segments and vowels. In the place of syllable boundary position features, there are measures for vowel/consonant switch boundaries. Features referencing lexical stress are omitted, as there is no reliable way to approximate them using only orthography. The set of these features and their descriptions is given in Table 2.3:

Feature names	Description	Number of distinct features values
length of Word 1/Word 2	number of characters in each input word	2
proportion of Word 1/Word 2 preserved	proportion of characters from each input word that are preserved	2
medial overlap	whether the prefix/suffix have overlapping character sequences	1
blend candidate vowel count	the number of vowels (syllable nuclei) in the blend candidate	3
consonant/vowel boundary of switchpoint (Word 1, Word 2)	whether switchpoint in each word occurs at a vowel-consonant or consonant-vowel transition point	4

Table 2.3: Orthographically-derived feature set

Section 2.9: Measures of Model Performance

The primary model comparison metric is the proportion of blends for which the model correctly predicts the winning output. This metric is standard to the previous two blend modeling approaches. Due to the fact that the selection component assigns only 1 member of the candidate set, rather than assigning a class to all blends, typical measures for classification tasks such as accuracy, precision, and recall are not considered.

In addition to prediction rate, each model and feature set pairing will be evaluated on Levenshtein edit distance between the predicted output and the correct blend form. This metric shows the number of character edits (additions, deletions, and substitutions) required to make two strings identical. It was also used by both Deri & Knight and Gangal et. al, and will therefore allow the performance of this model to be more directly compared to theirs.

The final metric used will be the number of blends for which the correct candidate is among the three candidates with the highest probability assigned by the model. This metric was not used in previous blend modeling approaches, but is introduced here to account for the fact that, for naturally occurring lexical blends, there are often many plausible candidates when the blend is initially coined (Arndt-Lappe & Plag). Ultimately, one form will win out, but before it has been lexicalized there will often be alternatives. For this reason, we consider the top 3 candidates to account for the possibility that the model has given a high probability to the correct candidate, but ultimately selected an alternative blend as the winner.

Section 2.10: Addressing Model Feature Correlations

Considering the fact that many of the features in the model are based on alternative linguistic analyses of blends, the feature system has many variables with high correlation - among the 25 features, 16 correlations with an absolute value of .75 or higher exist, as well as 34 correlations with absolute values between .5 and .75 and 32 correlations between .4 and .5, all of which are statistically significant.. As such, training on the whole feature set may lead to results with limited validity beyond these experiments, and may also limit the interpretability of the feature weights of the model, or even cause the model to fail to converge on the optimal weighting for any of the features. As such, efforts were made to reduce the feature set and reduce the potential risks of feature collinearity by manually selecting a subset of the feature matrix.

Features with correlations over 0.75 or multiple correlations over 0 .4 were removed from the feature set, starting with the features with the most correlations. In selecting between features with equally large correlation coefficients, the feature that was expected to be more linguistically informative was selected. This process resulted in two sets, differing only in one pair of features with roughly similar correlations with other model features. The first of these uses the proportion of word segments preserved as input features, the other uses the proportion of syllables preserved. In both of these subsets, all features based on the number of syllables between the switchpoint and the edge of the input words were removed, as well as the measures of initial input word length and whether the switchpoint is at a syllable nucleus boundary. This results in a feature set with no correlations with absolute values greater than .5, though 8 statistically significant correlations between .40 and .50 are still found among the remaining features of both models.

A fourth and final subset of the features was generated by only using the top 15 most informative constraints in the model. This algorithmically selected set was initially intended to be chosen using recursive feature elimination, but this failed to reduce the number of features, as all 24 were retained.

CHAPTER 3: RESULTS

Data for each corpus is given in tables which show value for proportion of blends correctly predicted, mean Levenshtein edit distance, and proportion of blends selected within the top 3 candidates, averaged across 10 folds. These averages are separated first by model type and secondarily by feature set.

For all corpora, the highest performance was achieved by using the full feature dataset, regardless of model type. For each corpus, 2-level ANOVA was performed on the prediction rates in order to compare the various model variations to one another, with the model type and feature subset type as the variable conditions, the distributions measured were the random samples produced during cross validation.. Finally, results from the orthographic feature models are given for each corpus, along with paired t-tests between the average performance of orthographic and phonological feature sets. Training and test folds of the data are held consistent across model and learner types to allow for a matched sample approach for this parametric test to compare across model types (Wong 2017).

Section 3.1: Shaw Corpus Results

The Shaw corpus was the primary corpus used for developing the model, and boasts the highest overall success rates in predicting blends, outperforming the results of the other corpora on all metrics and demonstrating large improvements over previous machine learning attempts at modeling blends. The performance of each feature set/learner type pairing with regards to percentage of correctly identified blends, percentage of blends found in the top 3 candidates, and Levenshtein edit distance is given in Table 3.1:

Model type	Feature set	Percent correct	In top 3	Mean edit dist.
LASSO regression	all available features	66.09%	83.46%	0.9258
	segmental features	64.17%	78.98%	0.9787
	syllable count features	60.25%	77.61%	1.1515
	auto top 15 features	64.17%	79.17%	1.0171
Ridge regression	all available features	66.10%	81.18%	0.9331
	segmental features	63.99%	78.52%	0.9778
	syllable count features	60.70%	76.70%	1.1222
	auto top 15 features	64.27%	78.89%	1.0108
Random Forest	all available features	73.96%	87.30%	0.5774
	segmental features	72.58%	86.47%	0.6453
	syllable count features	66.46%	82.91%	0.7769
	auto top 15 features	71.12%	86.74%	0.6472
Polynomial regression	all available features	74.13%	88.03%	0.5823
	segmental features	71.67%	85.65%	0.6571
	syllable count features	67.28%	83.00%	0.9105
	auto top 15 features	70.39%	85.10%	0.7511

Table 3.1: Shaw corpus model performance by learner type and feature set

For every model type, training using the whole feature system yielded the greatest accuracy and using the syllable only feature set had the lowest performance. Across all models and feature set pairings, the average percentage of correct predictions was 67.33%, with the correct candidate appearing in the top 3 an average of 82.48%, and the mean edit distance was .8540.

Two-way ANOVA results are given in the table below for the same corpus:

2-way ANOVA	Sum of Squares	Degrees of Freedom	*F* value	*p*-value
Model type	0.219	3	34.142	9.94E-17
Feature set	0.065	3	10.205	3.90E-06
Model type x feature set	0.009	9	0.470	0.893
Residual	0.308	144	-	-

Table 3.2: Two way ANOVA for Shaw corpus models

The ANOVA test demonstrates significant main effects for both model type and feature set, but no significant interaction effect between the two variables was observed. We also see the performance of the orthography only model on this corpus, with performance reported by feature set, as demonstrated in Table 3.3 and Table 3.4, which show the average performance of the models trained on phonological models and those using the purely orthographic feature set.

Orthographic Feature Models			
Model type	Percent correct	In top 3	Mean edit distance
LASSO regression	52.68%	67.54%	1.425
Ridge regression	52.68%	67.62%	1.427
Random Forest	54.90%	74.48%	1.106
Polynomial regression	57.27%	67.30%	1.170

Table 3.3: Orthographic model averages by learner type for Shaw corpus

Phonological Feature Models			
Model type	Percent correct	In top 3	Mean edit distance
LASSO regression	63.67%	79.80%	1.018
Ridge regression	63.76%	78.82%	1.011
Random Forest	71.03%	85.85%	0.662
Polynomial regression	70.87%	85.44%	0.725

Table 3.4: Phonological feature performance on Shaw corpus aggregated by learner type

The average correct predictions using only orthographic measures, such as number of characters deleted and number of vowels preserved, as proxies the phonological features of this feature set was 54.38%, with the candidate appearing in the top 3 candidates on average 69.23% of the time. The mean edit distance across all feature subsets was 1.2820. On all three performance metrics, the orthographic model has lower performance than the model using phonological features.

Section 3.2: Gangal et al. Corpus Results

The model iterations trained on the Gangal et al. corpus were far less successful than those trained on the Shaw corpus, but nevertheless showed an increase over previous benchmarks in the best cases, and roughly comparable performance to the benchmark in the leaner/feature pairings with the lowest metric scores. The table for the model performance metrics after training/testing on the Gangal et. al. corpus is given in the following table:

Model type	Feature set	Percent correct	In top 3	Mean edit dist.
LASSO regression	all available features	47.72%	68.74%	1.3628
	segmental features	46.17%	66.55%	1.4286
	syllable count features	43.42%	64.17%	1.5777
	auto top 15 features	46.16%	67.92%	1.4945
Ridge regression	all available features	47.72%	68.65%	1.3655
	segmental features	45.80%	66.91%	1.4478
	syllable count features	43.61%	65.09%	1.5603
	auto top 15 features	46.16%	67.91%	1.4826
Random Forest	all available features	58.14%	79.15%	0.8920
	segmental features	57.32%	67.54%	0.9513
	syllable count features	51.01%	75.78%	1.1460
	auto top 15 features	55.38%	77.42%	1.0131
Polynomial regression	all available features	59.51%	79.26%	0.8910
	segmental features	54.21%	75.60%	1.0649
	syllable count features	54.21%	75.05%	1.0950
	auto top 15 features	54.30%	77.43%	1.0636

Table 3.5: Gangal et. al. corpus model performance by model type and feature set

For this corpus, the mean percentage of times the correct blend was identified as the top candidate was 50.68%, with an average percentage of blends in the top 3 candidates of 71.45%. The average Levenshtein edit distance across all model and feature types was 1.2398. The best performing model/feature pairing with regards to all 3 metrics was logistic regression using polynomial features, training on the entire available feature set, with an average of 59.51% percent of blends identified as the first candidate, and an average of 79.26% of corrected candidates appearing in the top 3 choices of the model. The average edit distance for this model was 0.8910.

The previous benchmark for this corpus was 48.75% correct predictions, with a mean edit distance of 1.12. This was achieved by using an LSTM model that employed the exhaustive generation strategy and utilizing strategies such as attention word embeddings and model ensembling. Using only regression and low-order polynomial terms, this model outperformed that previous model on both metrics used in that study, with 59.51% of blends correctly predicted when using the full feature set. Random Forest learners using the same feature set also improved on the LSTM model, with 58.14% of blends correctly identified.

As with the model results on the Shaw corpus, a statistically significant main effect was found for both the variable of Model type and Feature set, but no significant interaction effect was discovered.

2-way ANOVA	Sum of Squares	Degrees of Freedom	*F* value	*p-value*
Model type	0.380168	3	64.484407	1.72E-26
Feature set	0.04716	3	7.999279	5.73E-05
Model type x feature set	0.017468	9	0.987639	0.4528784
Residual	0.282984	144	-	-

Table 3.6: Two way ANOVA for Gangal et. al. corpus models

Using only orthographic proxies for the phonological features, the mean percentage of correctly identified blends across all models training/testing on the Gangal et al corpus was 37.63%, with an average of 55.19% of all blends located in the model's top 3 predictions. The mean Levenshtein edit distance was 1.93. Comparing the average prediction rate of the phonological feature set by model, the mean of differences was 13.05 percentage points, which was found to be significant at *p* =.001.

Orthographic Feature Models			
Model type	Percent correct	In top 3	Mean edit distance
LASSO regression	35.01%	50.32%	2.157
Ridge regression	35.15%	50.32%	2.157
Random Forest	40.04%	61.59%	1.631
Polynomial regression	40.32%	0.5853860753	1.756

Table 3.7: Orthographic model averages by model type for Gangal et. al. corpus

Phonological Feature Models			
Model type	Percent correct	In top 3	Mean edit distance
LASSO regression	45.87%	66.84%	1.466
Ridge regression	45.82%	67.14%	1.464
Random Forest	55.46%	74.97%	1.001
Polynomial regression	55.56%	76.83%	1.029

Table 3.8: Phonological feature performance on Gangal et. al. corpus aggregated by model type

Section 3.3: Deri & Knight Corpus Results

Results for all model/feature combinations using the Deri corpus are given in Table 3.9 on the following page.

Model type	Feature set	Percent correct	In top 3	Mean edit dist.
LASSO regression	all available features	56.13%	74.84%	1.0465
	segmental features	55.21%	73.60%	1.0866
	syllable count features	53.66%	69.92%	1.1795
	auto top 15 features	55.52%	76.35%	1.0521
Ridge regression	all available features	55.82%	74.84%	1.0586
	segmental features	56.44%	72.38%	1.0610
	syllable count features	53.96%	69.61%	1.1916
	auto top 15 features	54.91%	80.35%	0.8287
Random Forest	all available features	60.76%	79.63%	0.8131
	segmental features	60.39%	79.79%	0.8313
	syllable count features	56.42%	75.79%	0.8924
	auto top 15 features	59.15%	80.35%	0.8242
Polynomial regression	all available features	64.42%	84.02%	0.7232
	segmental features	63.83%	80.35%	0.7910
	syllable count features	59.21%	80.69%	0.8802
	auto top 15 features	62.30%	81.89%	0.7801

Table 3.9: Deri & Knight corpus model performance by model type and feature set

For this corpus, all models trained using the novel system perform with higher prediction rates than the corpus benchmark. The previous standard set by Deri & Knight was 48.75%, and the lowest performing model and feature set in this test was LASSO regression using syllable-based features, which had a prediction rate of 53.66%. The mean for this dataset was 58.01%, with a mean of 77.15% of candidates correctly predicted within the first three selections.

Only 2 models in this system had a lower average Levenshtein distance than the benchmark of 1.12, and the average Levenshtein distance on the corpus as a whole was .94. The highest performing model on this corpus (once again polynomial regression) achieved a prediction rate of 64.42% ,more than 15 percentage points above the baseline.

2-way ANOVA	Sum of Squares	Degrees of Freedom	*F* value	*p*-value
Model type	0.147	3	9.432	1.00E-05
Feature set	0.029	3	1.889	0.134
Model type x feature set	0.005	9	0.112	0.999
Residual	0.749	144	-	-

Table 3.10: Two way ANOVA for Shaw corpus models

As with the previous corpora, models were trained using the orthographic feature set to compare the utility of this set to the proposed phonological features. The performance of these orthographic feature-trained model iterations is given and compared to models with the same learner type, trained and evaluated over the same test folds using phonological features. This comparison once again shows an apparent increase in performance for models trained using orthographically-derived, non-phonological features.

Orthographic Feature Models			
Model type	Percent correct	In top 3	Mean edit distance
LASSO regression	45.12%	60.32%	1.539
Ridge regression	44.84%	60.59%	1.533
Random Forest	48.91%	72.25%	1.160
Polynomial regression	51.07%	75.79%	1.155

Table 3.11: Orthographic model averages by model type for Deri & Knight corpus

Phonological Feature Models			
Model type	Percent correct	In top 3	Mean edit distance
LASSO regression	55.13%	73.68%	1.091
Ridge regression	55.28%	74.29%	1.035
Random Forest	59.18%	78.89%	0.840
Polynomial regression	62.44%	81.74%	0.794

Table 3.12: Phonological feature performance on Deri & Knight corpus aggregated by model type

The average prediction rate for all orthographic models was 47.49%. When compared with phonological features using the same models and same data training/test splits, the mean difference in prediction rate was 12.95 percentage points, which was found to be significant at *p* = 8.94E-04. The average of 67.24% of correct candidates appearing in the top 3 predictions of all models. The mean Levenshtein distance was 1.3467.

In order to verify that the difference in performance between the orthographic and phonological models was not due merely to the exclusion of suprasegmental features in the orthographic case, a new set of trials was performed using a subset of the phonological features which excluded all features that referenced prosodic information that would be unavailable at the orthographic level. This resulted in a feature set with a total of 16 terms, slightly larger than the total of 12 terms used by the orthographic models. The average percentage of accurate predictions was again compared to the performance corresponding learner types for the orthography-only feature sets.

	Shaw	Gangal et al.	Deri & Knight
LASSO	12.58	11.61	10.15
Ridge	11.94	11.20	11.04
RF	17.50	15.55	11.23
Polynomial	15.95	23.47	11.88
Average	14.492	15.46	11.08
p-value	0.00167	0.01224	0.00007

Table 3.13: Average difference in percentage points between models trained with non-prosodic phonological features and those trained with orthographic features, given by learner type and corpus

A two-sample *t*-test of the differences in accuracy between the models trained using non-prosodic phonological features and those trained with orthographic features revealed a significant difference between the two feature sets for all corpora. This test suggests that the differences in prediction rates between the phonological models and orthographic models is not due solely to the inclusion of suprasegmental features.

Section 3.4: Resulting Coefficients

Feature coefficients for the LASSO regression using the full set of available features and the manually selected subset with segmental measures are presented, along with the top 25 largest feature weights by magnitude for the polynomial regression using all features and segmental features. This model showed the better performance of the two regression models which only had the unaugmented feature set.

The feature coefficient weights learned by this iteration of the model are given in the following table, sorted in descending order of the absolute value of the coefficient.

Feature name	Coefficient value
Word 1 proportion surviving segments	4.558
Word 2 proportion surviving segments	2.998
Candidate has medial overlap	2.053
Switchpoint at Word 1 syllable nucleus boundary	0.897
Switchpoint at Word 1 onset boundary	0.861
Word 2 primary stress syllable survived	0.827
Candidate length (syllables)	-0.697
Switchpoint at Word 1 coda boundary	0.593
Word 1 length	0.507
Distance from Word 1 left edge to switchpoint	-0.455
Distance from Word 2 right edge to switchpoint	0.453
Switchpoint at Word 2 syllable nucleus boundary	-0.437
Word 2 proportion surviving syllables	0.400
Word 1 primary stress syllable survived	0.306
Switchpoint occurred at point of Word 2 primary stress	0.293
Switchpoint at Word 2 coda boundary	-0.223
Switchpoint at Word 2 onset boundary	0.165
Blick phonotactic learner score	-0.148
Distance from Word 2 right edge to primary stress	-0.106
Distance from Word 1 right edge to primary stress	-0.089
Word 1 proportion surviving syllables	0.084
Distance from Word 2 left edge to primary stress	-0.082
Distance from Word 1 left edge to primary stress	-0.064
Word 2 length	0.035

Table 3.14: Coefficient values for LASSO regression model with full feature set

The most strongly weighted features are the proportion of segments retained from each of the input words, followed closely by whether or not a blend candidate has medial overlap between the phonemes. No other feature was nearly as strongly weighted in the regression. No features were fully eliminated from the model, though this may be a result of averaging the weights after cross-validation.

One surprising finding is that the BLICK score was among the 10 features with the lowest feature weights. This feature was deliberately selected in order to limit the possibility of ill-formed candidates being selected as the winning output, yet seems to be doing very little to influence the model.

The feature weights learned by the LASSO regression learner using the segmentally-based manual subset features were also examined:

Feature name	Coefficient value
Word 2 proportion surviving segments	5.516
Word 1 proportion surviving segments	3.630
Candidate has medial overlap	2.128
Candidate length (syllables)	-0.670
Switchpoint occurred at point of Word 2 primary stress	0.445
Distance from Word 2 left edge to primary stress	0.373
Switchpoint at Word 2 coda boundary	-0.296
Switchpoint at Word 1 onset boundary	0.218
Switchpoint at Word 2 onset boundary	0.195
Distance from Word 1 right edge to primary stress	0.174
Distance from Word 1 left edge to primary stress	0.163
BLICK phonotactic learner score	-0.132
Distance from Word 2 right edge to primary stress	0.126
Switchpoint at Word 1 coda boundary	-0.042

Table 3.15: Coefficient values for LASSO regression model with segmental subset features

The relative importance of coefficients in this feature set is nearly identical to those found using the full feature set - the proportion of segments preserved from both words and whether the blend has overlap are once again the most highly weighted features, while BLICK score and syllable constituent features were again given low coefficient values.

Another interesting pattern which emerges across the two datasets is the tendency for features relating to Word 1 to have higher coefficient values than those relating to Word 2. This is somewhat surprising, given that most researchers find that the length of Word 2 and the proportion preserved from that word should be more informative in determining switchpoint.

The 25 coefficients from the Polynomial regression model with all features that had the greatest absolute value are given in Table 3.16.

Feature name	Coefficient value
W1 length * W1 proportion surviving syllables	5.171
W1 proportion surviving syllables	-4.954
Distance from W1 left edge to switchpoint	-4.260
W1 prop. segments * W2 prop. syllables	2.994
W2 prop. syllables * Has medial overlap	2.924
W1 prop. syllables * W1 prop. segments	2.082
W2 length* W1 prop. syllable	1.723
W2 prop. syllables * Switchpoint at W1 onset boundary	1.678
W2 prop. syllables * Switchpoint at W2 onset boundary	1.657
W1 prop. segments * overlap	1.603
W2 prop. syllables * W1 distance right edge to primary stress	1.588
W2 prop. syllables * W2 distance left edge to primary stress	-1.496
W1 prop. syllables * W2 prop. syllables	-1.480
W1 length * W2 prop. segments	1.460
W2 prop. segments * W2 distance left edge to primary stress	-1.422
W2 prop. segments * W2 primary stress preserved	1.370
Has medial overlap * Switchpoint at W2 nucleus boundary	-1.363
Switchpoint at W1 onset boundary	-1.327
1	-1.214
Switch at W1 nucleus boundary * W2 primary stress survived	1.190
Switch at W1 nucleus boundary * W1 primary stress survived	-1.174
W1 proportion segments * W1 distance right edge to primary stress	-1.150
W2 proportion segments * Switchpoint at W1 onset boundary	-1.074
Candidate length (syllables) * W2 primary stress preserved	1.056
Switch at W2 onset boundary * W2 primary stress preserved	1.031

Table 3.16: Coefficient values for Polynomial regression model with all available features

Reviewing these coefficient weights reveals that interaction terms account for nearly all of the most important features in the model, with only two x^1-level features and no polynomial terms appearing in the top 25 (though the x^0 term does appear in the top 25). These weights also demonstrate an increased importance of syllable-based measures of proportions preserved from each input word in the Polynomial model when compared to regular LASSO regression. The model also weights features based on input word stress much higher than the regular LASSO models weight stress features alone, which may be reflective of the findings of Arndt-Lappe & Plag (2013) that metrical properties are among the most important factors in determining blend switchpoints and structure.

On the other hand, many of the interaction terms with the greatest absolute feature weights are *nearly* identical measures of blend survival, even though they are not truly the same feature. For example, ```W1 prop. syllables * W1 prop. segments``` measure nearly identical variables of the candidates, but the model assigns a stronger weight to their interaction term than the squared value of either term. This may account for the difference in performance between the full-feature models and those which used only a subset. It seems that including subtly different features allows the model to better represent the actual distributions of blends.

Notably, of the 25 interaction features with the strongest weight values, 14 of them include at least one binary feature value. Of these, all but 3 are given positive feature weights. This seems to suggest that certain features contribute positively to the likelihood of valid blend formation only conditionally. For example, the interaction term 'proportion of Word 2 syllables preserved * has medial overlap* is given the weight 2.924, indicating that an increase in the proportion of syllables preserved from the 2nd word corresponds to an increased likelihood of a candidate as a valid blend, but only if the blend candidate has an overlap in phonetic material between the two input words. Similarly, the positive weight associated with the interaction term ```Switch at W2 onset boundary * W2 primary stress preserved``` indicated that the greater the number of syllables in a blend candidate, the more probable it is as the desired output, but only as long as the primary stress of Word 2 is preserved.

Feature name	Coefficient value
W2 prop. syllables * Has medial overlap	2.924
W2 prop. syllables * Switchpoint at W1 onset boundary	1.678
W2 prop. syllables * Switchpoint at W2 onset boundary	1.657
W1 prop. segments * overlap	1.603
Has medial overlap * Switchpoint at W2 nucleus boundary	-1.363
Switchpoint at W1 onset boundary	-1.327
Switch at W1 nucleus boundary * W2 primary stress survived	1.190
Switch at W1 nucleus boundary * W1 primary stress survived	-1.174
W2 proportion segments * Switchpoint at W1 onset boundary	-1.074
Candidate length (syllables) * W2 primary stress preserved	1.056
Switch at W2 onset boundary * W2 primary stress preserved	1.031

Table 3.17 Coefficient values from interaction terms comprised of at least one binary-valued model feature using Polynomial Regression with full feature set

Of the 325 features of this model, 84 were assigned a weight of 0 in every k-fold iteration and were effectively removed from consideration by the model. An additional 41 terms received an average coefficient value of less than 0.01 Most of these are features with small coefficient weights in the LASSO regression models and polynomial or interaction terms dependent on such terms, with the exception of the feature '*Has medial overlap*', which was assigned the third highest weight in both LASSO models, yet was effectively eliminated from the Polynomial model.

Because syllable proportion features and their interaction terms were so highly weighted in the Polynomial regression model using the full feature set, the top 25 most weighted features for the Polynomial regression using segmental proportions but not syllable proportions is drastically different from the model that utilizes all available features. These values are given in Table 3.18 on the following page:

Feature name	Coefficient value
W1 prop. segments * W2 prop. segments	7.897
W2 prop. segments^2	-2.171
W2 prop. segments * Has medial overlap	1.786
W2 prop. segments * W1 distance right edge to primary stress	1.777
W1 distance right edge to primary stress	-1.570
Switch at W1 onset boundary	1.507
Switch at W1 onset boundary^2	1.459
W2 prop. segments * Switch at W2 onset boundary	-1.441
W1 prop. segments * Has medial overlap	1.341
1	-1.318
W1 prop. segments * W1 distance right edge to primary stress	-1.305
W1 prop. segments * Switch at W2 primary stress	1.288
Switch at W1 onset boundary * Switch at W2 onset boundary	-1.082
W1 prop. segments * Switch at W2 onset boundary	1.050
Has medial overlap * Switch at W2 primary stress	-1.049
Has medial overlap * Switch at W1 onset boundary	-1.032
W2 distance left edge to primary stress	-0.979
W1 distance left edge to primary stress	-0.935
W2 prop. segments W1 distance left edge to primary stress	0.933
Switch at W1 onset boundary * W1 distance right edge to primary stress	-0.881
Candidate length (syllables) * W2 distance right edge to primary stress	0.827
Candidate length (syllables) * W2 distance left edge to primary stress	0.823
Switch at W1 coda boundary * Switch at W2 primary stress	-0.823
Switch at W2 coda boundary * W2 distance left edge to primary stress	0.819
Switch at W2 primary stress ^2	0.780

Table 3.18: Coefficient values for Polynomial regression model with manual subset features

The strongest coefficient weights for this model include many more x^2 terms and x^1 terms, though interaction terms still comprise most of the list. Somewhat surprisingly, terms based on the proportion of preserved segments account for a greater number of terms in this set, with 10 of the highest weighted features being terms based on the proportion of segments preserved in one of the input words, compared with 6 such features when using the full dataset. In this feature set, 21 terms were consistently weighted at 0, and 24 additional terms received an average weight of less than 0.01 across random instances of the model. The feature '*has medial overlap'* was not among these features, again being assigned a low feature weight at just 0.062, but several of its interaction terms were once again among the strongest constraint weights. This again indicates the usefulness of such binary variables in assigning conditional weights to features.

The overall trend revealed by the feature weighting in many ways demonstrates the core conundrum of blend formation, which is to preserve enough material from each word but not too much. This can be seen by the way that variables like 'W1 proportion surviving syllables' and interaction terms derived from it have seemingly conflicting values across different learner types and feature sets, being sometimes assigned positive coefficients and sometimes negative ones. In the full feature set Polynomial learner trial, for example, the term itself is given the weight -4.954 while the interaction term ```W1 length * W1 proportion surviving syllables ``` is given a weight of 5.171, and these two features are the two strongest feature weights learned by the model.

While initially seeming contradictory, this shows that the model is balancing how much of Word 1 to preserve by assigning a lower probability to candidates which preserve more of Word 1, but for blends with longer Word 1 inputs, it is giving preference to longer candidates. In this way, the different feature sets and learning components used seem to agree on what the most important features are (proportion of syllables/segments preserved, medial overlap, etc.) but the precise strategies used to make predictions based on these features and the pattern of feature coefficients learned varies greatly depending on the other available features and the complexity of the model's learner.

Section 3.5: Output Candidate Error Analysis

The chosen output forms of the best performing model were analyzed to determine the most common failings of the model and to determine the plausibility of the predicted forms as words of English. This analysis demonstrated that incorrect blend predictions are typically one single syllable longer than the desired output, having preserved one more syllable than desired from one of the input words. Notably, this shows a failure of the model to capture the generalization that blends tend to have the same number of syllables as one of the input words, most generally Word 2. Less frequently, some incorrectly chosen candidates were produced when the switchpoint was misplaced by simply preserving one segment too many from one word and deleting one too many from the other, resulting in blends that missed the target output with a difference of only a single segmental substitution.

Word1	Word2	Desired Output	Model Prediction
Europe	Asian	Eurasian	Europasian
Dixie	Democrats	Dixiecrats	Dixiemocrats
boy	burlesque	boylesque	boyurlesque
potato	tomato	pomato	potatomato
fog	drizzle	fozzle	fogrizzle
line	trunk	lunk	linerunk
recollect	remember	recomember	recollember
bizarre	exotic	bizotic	bexotic

Table 3.19: Comparison of predicted output and target blend forms for example blend word pairs

3.5.1: Error Analysis

A random sample of 100 incorrectly predicted blend outputs was drawn from the blends misclassified in the trial with the highest performing model/feature pairing (Polynomial regression with full features using Shaw corpus) to analyze the patterns of misclassification. Since the specific task of the model architecture is switchpoint selection, or determining how much of each word to preserve, these incorrect predictions were analyzed in terms of whether the candidate preserved too much segmental material from a particular input word, or deleted too much.. Candidates could have errors for one or both words, but could only have one type of error (over-preserved or under-preserved) for each input word.

Of all incorrect predictions in the sample, 91 could be described in terms of misplaced switchpoints. The total number of over-preservation and under-preservation from each input word was recorded, and the total count of each error type is given in Table 3.18 .

Word 1 proportion preserved too great	Word 2 proportion preserved too great	Word 1 proportion preserved too small	Word 2 proportion preserved too small
40	33	15	18

Table 3.20: Switchpoint error count for incorrect blend predictions

Incorrect blends were found to retain too much phonetic material from input words far more often than they retained too little, and seemed to over-preserve segmental material from Word 1 far more often that under-preserving from Word 1. Furthermore, Word 1 is under-preserved less often than Word 2. This demonstrates that, in addition to its failure to match blend length to input word length, the model is also exhibiting a tendency to preserve too much material from Word 1.

Given this pattern among the incorrect predictions, a possible improvement to the model may be to include additional model features which specifically measure the relationship between the candidate length and the length of the input words. This could be parameterized in terms of difference in length, but could also be given in binary variables (whether the candidate matches the input words in number of syllables), as such binary variables were shown to be useful interaction terms to allow the model to give conditional weights to other features.

Additional categorical variables to represent the part of speech of each input word could be useful, as it has been noted that the syntactic and semantic role of the input words in the blend can influence the proportion of each word preserved in the final blend (Shaw et al. 2014) . These may be particularly useful features to the model architecture as they may provide the ability for more conditional weighting of the most important features, similar to the existing binary features like the feature `*has medial overlap*`.

The remaining 9 candidates were results of errors in the candidate generation process. Five of these were incorrectly tagged, and did in fact produce the correct phonological output, but the label was misassigned due to mismatches in orthography. For example, the desired blend output for the pairing of ‘Europe’ with ‘bureaucracy’ is ‘Eurocracy’, but the model gave the output ‘Eureaucracy.’ The remainder were mismatched only because of an incorrectly predicted vowel reduction, or preserved post-syllabic [ɹ] when it was expected not to be preserved. For the incorrectly labeled candidates as well as the candidate with unexpected vowel reduction, an improvement feature extraction and labeling steps during candidate generation process could resolve these errors completely.

In order to compare the well-formedness of these incorrect blend choices, BLICK scores were calculated for a random sample of 80 incorrectly predicted blend outputs and 80 random words of English, sampled from the CMU pronouncing dictionary. The resulting average BLICK score for the random words was 4.24, with standard deviation 4.14. The predicted blend forms had a mean BLICK score of 6.48, with a standard deviation of 4.57. These values are compared in the chart below:

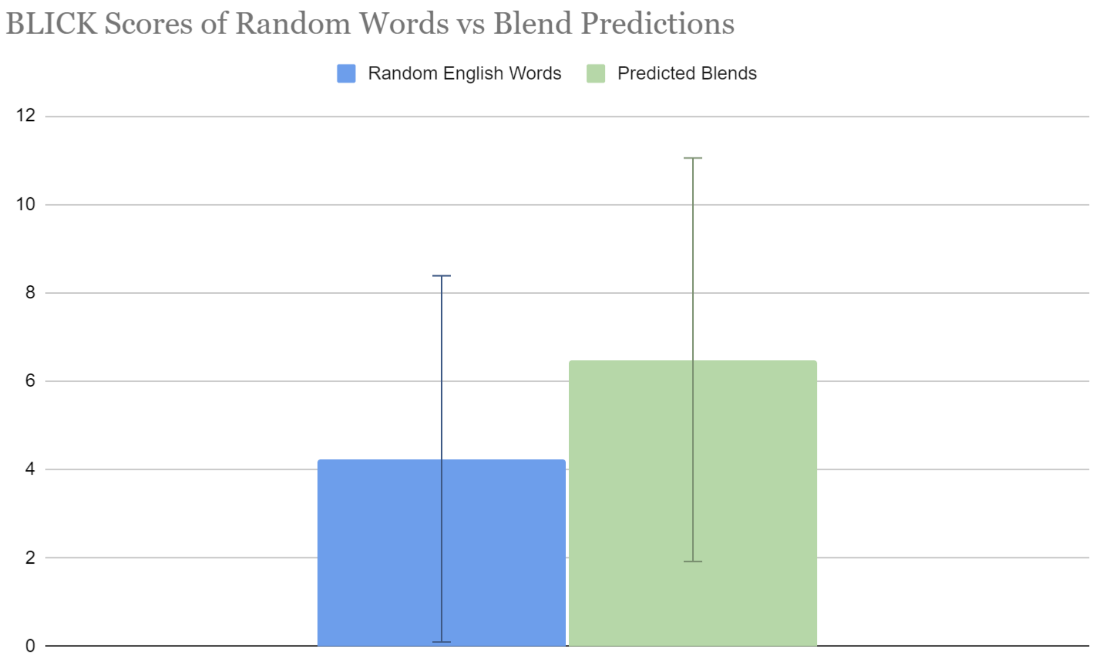


Figure 3.1: BLICK scores for random CMU dictionary words vs incorrect blend predictions

One final observation about qualitative properties of the model predictions is that some of the incorrect candidates chosen by the model are potentially better blends of English than the actual desired output. For example, the model predicts that the ideal blend of ‘female’ and ‘macho’ should be ‘femacho’ rather than ‘facho’, and also chooses ‘smuffocate’ instead of ‘smothercate’ as the blend form of ‘smother + suffocate’, both of which are judged by the author to be highly plausible alternatives to the actual desired output . It is possible that experimental trials with other human graders could show that the model is in fact learning to create valid blends, but simply chooses some candidates which were not ultimately selected as the lexicalized form for a particular input word pair.

CHAPTER 4: DISCUSSION

Section 4.1: Comparison to Previous Blending Models

It is clear that the methodology described in this project can improve the accuracy of predictions of lexical blends of English compared to previous data-driven models. Using the Deri & Knight corpus, the best performing model in this study (logistic regression with polynomial features, using full data set) correctly identified an average of 64.42% of blends in English, compared with a maximum of 45.39% for the highest performing cross-validated model result reported in Deri & Knight's original paper. In fact, even the model/feature pairing with the lowest performance for the new model on this corpus (LASSO regression with manually selected features) still achieved an average correct blend prediction rate of 53.66%, and the average for all models on this corpus was 58.01%.

Levenshtein distance was similarly improved when training/testing on this corpus, with an average edit distance between the predicted output and desired output of .9400 and a low of only .7232, compared to an edit distance of 1.59 reported for Deri & Knight's best performing model.

For the Gangal et. al. corpus, the difference in performance of the current proposed architecture compared to that of the benchmark model from the original authors is less drastic, but the best iterations of the present model still show a notable increase in the mean proportion of correctly identified blends and decrease in Levenshtein distance when compared with the best iterations of the present model.

Section 4.2: Differences Between Feature Sets

The difference in performance between the models trained on phonemic and suprasegmental features and the models using only orthographic proxies for these features can provide additional evidence to this point. Since the models using phonological features had significantly greater prediction rates for blends, we have reason to believe that including phonological features, rather than simply character-based measures, does improve model performance. On the other hand, the fact that, for the Deri & Knight corpus, the orthographic models using polynomial regression outperformed the previous benchmark is evidence for the effectiveness of the overall model architecture and the use of the exhaustive generation approach.

Section 4.3: Differences Between Model Variations

The consistently high performance of polynomial features seemed initially to be due to the increased ability for the model to handle nonlinearities in the feature space, rather than the predictive power of new features alone. Logistic regression models are limited by the assumption of linearly separable decision boundaries in the feature space, and since the features used in the model are almost all violable to some extent, there is likely to be significant overlap in many, if not all of the distributions features used to distinguish valid blends from invalid candidates.

Polynomial regression is known to be better able to approximate a wider range of input variable distributions than base logistic regression, and so seemed likely to be the root cause of the increased performance. When the Polynomial regression with all features was limited to the top 25 features with the greatest R^2 value, the prediction rate fell to 61.98%, down from 74.13% when using all available features, suggesting that the polynomial features are indeed better able to approximate the distribution due to increased features with polynomial distributions, rather than simply utilizing features which are in themselves more informative.

The improved prediction rates observed in the random forest models also seem to support the hypothesis that nonlinearity is the principal reason for the differences in model performance, as random forest models have been shown to outperform regression models on data with known non-linearities, and are not limited by assumptions of linear separability in the way that regression models are (Rigatti 2017).

However, the fact that the actual learned coefficients with the greatest absolute values for the polynomial regression were nearly all interaction terms and the fact that 125 of the 325 features were given average weights with an absolute value less than 0.01 may indicate that, in reality, it is the inclusion of interaction terms which is driving the increase in performance of this model. The specific interaction terms which are selected by the model are also more in line with the hypotheses about blends offered up by Arndt-Lappe & Plag (2013), namely that prosodic structure is a key component in predicting how words will be combined into blends. Future analysis on this model architecture should therefore include measures to better determine the importance of these interaction terms in predicting lexical blend output forms.

Section 4.4: Differences in Model Performance by Corpus

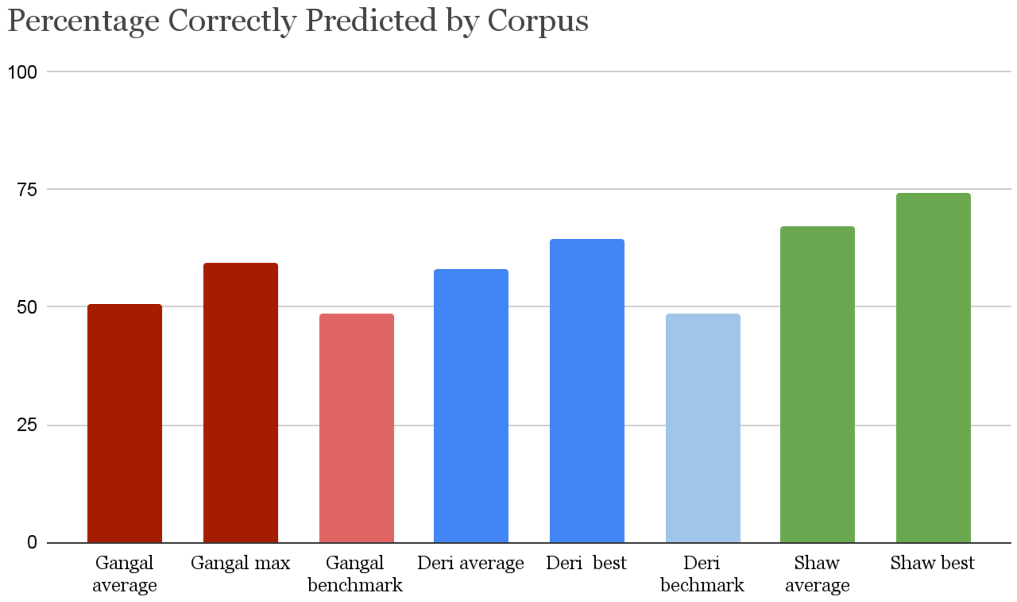


Figure 4.1: Maximum and average prediction rates for each model, sorted by corpus

The apparent discrepancies in performances of the models between blend corpora could present a potential area of concern for the validity of the models trained, but these differences are most likely a result of the varying sizes, as well as substantial qualitative differences of the corpora which were examined. As stated previously, the data in the Gangal et. al. corpus was not cited in any linguistic publication, but rather was obtained through reference to online forums. This may mean that this corpus contains a larger number of novel blends, rather than those that have actually entered into use in the lexicon of a particular speech community. It also may be the case that this model is simply more suited to learning lexicalized blends compared to novel ones.

The Deri & Knight corpus seems more similar to the Shaw corpus qualitatively, but is also highly restrictive in the blends it considers and is quite small - during the training stage using this corpus, the full size of each training set was only 293 instances, despite training on

90% of the dataset. It is plausible, then, that the difference in performance between the Deri & Knight and Shaw corpora is simply due to the Shaw corpus being able to use more examples to establish a more accurate feature weighting.

Section 4.5: Implications of Observed Feature Weights

The low coefficient weight of the BLICK score seems to suggest that accounting for phonotactic probability is less important for blend prediction than accounting for factors like the proportion of each word saved and whether a blend has medial overlap. If this limitation on phonotactic features as a predictor extends to orthotactics, this may offer some explanation for the limited success of the Gangal et al model, as the use of character embeddings rather than actually considering the number of preserved characters directly may have affect the model's ability to learn the relevant patterns that are actually most important in predicting blend structure.

Despite the fact that this feature seems to contribute little to the probability of a sequence as a valid blend, yet the models' mean prediction rates and edit distances both improve drastically using the feature set outlined in this model, and the incorrectly predicted forms are not judged to be especially ungrammatical by the BLICK scorer. This seems to suggest the combination of other features in the model are more helpful in leading to valid outputs than accounting for phoneme sequences. This could explain the limited success of Deri & Knight's approach, as it relied almost entirely on phoneme sequence probabilities to train the model. To a lesser extent, this may also apply to Gangal et. al., as they used character-based vector embeddings to train their model.

CHAPTER 5: CONCLUSION

Section 5.1: Summary of Findings

The model architecture described in this paper has shown that there may still be areas of research in the modern natural language processing (NLP) landscape in which observable, linguistically-derived features may still be useful, and can in fact outperform complex and sophisticated deep learning methods, with a fraction of the model complexity.

Using very minimal data manipulation in the form of 2nd order polynomial and interaction terms, this model outperformed the LSTM's baseline across all feature sets. Although this approach requires a notable amount of feature processing, it can yield results that are worth the effort of hand-chosen feature extraction.

By comparing to orthographic models, the usefulness of phonological features and their importance to the model architecture has been shown. The features significantly improve the performance of the architecture over using orthographic features alone. Even when the models fail, they typically generate phonotactically plausible candidates with orthographic sequences that are similar to the desired outputs, as shown by the low BLICK markedness scores of incorrect candidates and the low average Levenshtein edit distances of the output forms.

With some improvements in the candidate selection and feature processing components of this model, the architecture could be applied to theoretical linguistic questions about the grammar of blends, as this architecture is not only more effective and simpler than other models but is more transparent, with more interpretable features.

Section 5.2: Applications for Future Research

The next most obvious test to validate this model architecture and feature set would be to apply it to a task of novel blend generation and compare its predictions to those generated by speakers of English. This would provide insight into whether the discrepancy in performance between the different corpora is likely to be due to difference in stages of lexicalization. This would also allow the reliability of the model to be tested, and could provide insight into the features used in making new blends and the relationship between the highest probability candidates the model predicts compared to the frequency of possible novel forms speakers create.

More interestingly for the model as a whole, it would allow the model to be compared to a modern large language model to test blend generation abilities. Since large language models have access to so much data, it would not be possible to rank their performance on blend prediction for lexicalized blends, but with novel human utterances, a large model could be compared to the model architecture laid out in this thesis.

Another useful comparison would be to train an LSTM using the phonologically-derived features of this model, rather than the character embeddings used by Gangal et al. This would allow the predictive power of the feature set to be more directly compared with those created by character-based attention embeddings. This could also provide additional data on the hypothesis that improved performance is correlated with increased ability to model non-linear data, as deep neural nets have been shown to improve performance over random forest in other domains (Ahmad., Mourshed, & Rezgui, 2017). This could also lead to a further increase in prediction accuracy and decrease in Levenshtein distance.

One important advantage that this model architecture has over models with features obtained through unsupervised learning is the ability it provides to directly manipulate the feature space. This characteristic of the model, coupled with the relative ease of training offered by the reduced learner complexity, gives this architecture the ability to easily be modified by small changes in the feature space in order to both maximize the predictive power through additional feature engineering and to empirically examine linguistic hypotheses of blend formation.

By more carefully crafting the input features, the model could be used to test various assumptions about blends, such as the ways that input word length and blend length interact. This question has been debated thoroughly in the literature, but a predictive model using the architecture laid out here could be useful in determining the relative usefulness of different length constraints in predicting actual blend structures.

This architecture could similarly be used as a testing ground for new proposed features used to describe blends. Variables from other domains, such as word recognition point from psychology or perceived similarity to other words in the lexicon might also inform blend formation. One method of examining the usefulness of such features could be to include them in this model and to test the performance of the model after their inclusion.

REFERENCES

Ahmad, M. W., Mourshed, M., & Rezgui, Y. 2017. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. In *Energy and buildings*, 147. (pp. 77-89).

Arndt-Lappe, S., & Plag, I. 2013. The role of prosodic structure in the formation of English blends. *In English Language & Linguistics*, 17(3). (pp. 537-563).

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., & Hulden, M. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 10-22).

Deri, A., & Knight, K. 2015. How to make a frenemy: Multitape FSTs for portmanteau generation. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 206-210).

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. 2008. LIBLINEAR: A library for large linear classification. *In Journal of Machine Learning Research*, (pp. 1871–1874).

Gangal, V., Jhamtani, H., Neubig, G., Hovy, E., & Nyberg, E. 2017. Charmanteau: Character embedding models for portmanteau creation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2917-2922).

Gries, S. T. 2004. Shouldn't it be breakfunch? A quantitative analysis of blend structure in English. In *Linguistics*. 42(3). (pp. 639-667).

Gries, S. T. 2006. Cognitive determinants of subtractive word formation: A corpus-based perspective. In *Cognitive Linguistics*, 17(4).

Gries, S. T. 2012. Quantitative corpus data on blend formation: Psycho-and cognitive-linguistic perspectives. In *Cross-disciplinary perspectives on lexical blending*, 252, 145.

Hayes, B. 2012. BLICK: a phonotactic probability calculator (manual).

Hayes, B., & Wilson, C. 2008. A maximum entropy model of phonotactics and phonotactic learning. Linguistic inquiry, 39(3), 379-440.

Kelly, M. H. 1998. To “brunch” or to “brench”: Some aspects of blend structure. In *Linguistics,* 36(3).

Kubozono, H. 1990. Phonological constraints on blending in English as a case for phonology-morphology interface. *Yearbook of morphology*, 3, (pp. 1-20).

Kulkarni, V., & Wang, W. Y. 2018. Simple models for word formation in slang. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1424-1434).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. In *The Journal of machine Learning research*, 12, (pp. 2825-2830).

Rigatti, S. J. 2017. Random forest. In *Journal of Insurance Medicine*, *47*(1), (pp. 31-39).

Shaw, K. E., White, A. M., Moreton, E., & Monrose, F. 2014. Emergent faithfulness to morphological and semantic heads in lexical blends. In *Proceedings of the annual meetings on phonology* 1(1).

Thurner, Dick. 1993. Portmanteau dictionary: Blend words in the English language, including trademarks and brand names. Jefferson, NC: McFarland & Co.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *In Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1), (pp. 267-288).

Von Kleist, Adolf . 2015. Phonetisaurus G2P. GitHub repository github.com/AdolfVonKleist/Phonetisaurus

Wong, T. T. 2017. Parametric methods for comparing the performance of two classification algorithms evaluated by k-fold cross validation on multiple data sets. In *Pattern Recognition*, 65. (pp. 97-107).

Zoph, B., & Knight, K. 2016. Multi-Source Neural Translation. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 30-34).